SYMMETRIZED NEAREST NEIGHBOR REGRESSION ESTIMATES

R. J. Carroll[1]

W. Härdle[2]

[1] Department of Statistics, Texas A&M University, College Station, TX 77843 (USA). Research supported by the Air Force Office of Scientific Research and by Sonderforschungsbereich 303, Universität Bonn.

[2] Universität Bonn, Rechts- und Staatswissenschaftliche Fakultät, Wirtschaftstheoretische Abteilung II, Adenauerallee 24-26, D-5300 Bonn, Federal Republic of Germany.

ABSTRACT

We consider univariate nonparametric regression. Two standard nonparametric regression function estimates are kernel estimates and nearest neighbor estimates. Mack (1981) noted that both methods can be defined with respect to a kernel or weighting function, and that for a given kernel and a suitable choice of bandwidth, the optimal mean squared error is the same asymptotically for kernel and nearest neighbor estimates. Yang (1981) defined a new type of nearest neighbor regression estimate using the empirical distribution function of the predictors to define the window over which to average. This has the effect of forcing the number of neighbors to be the same both above and below the value of the predictor of interest; we call these symmetrized nearest neighbor estimates. The estimate is a kernel regression estimate with "predictors" given by the empirical disribution function of the true predictors. We show that for estimating the regression function at a point, the optimum mean squared error of this estimate differs from that of the optimum mean squared error for kernel and ordinary nearest neighbor estimates. No estimate dominates the others. They are asymptotically equivalent with respect to mean squared error if one is estimating the regression function at a mode of the predictor.

Key Words and Phrases: Nonparametric regression, kernel regression, nearest neighbor regression, bias, mean squared error.

Section 1: Introduction

We consider nonparametric regression with a random univariate predictor. Let (X, Y) be a bivariate random variable with joint distribution H, and denote the regression function of Y on X by $m(x) = E(Y \mid X = x)$. If it exists, let f_x denote the marginal density of X. A sample of size n is taken, (y_i, x_i) for $i = 1, ..., n$. Two common estimates of the regression function are the Nadaraya-Watson kernel estimate and the nearest neighbor estimate, see Nadaraya (1964), Watson (1964) and Stute (1984) for the former, and Mack (1981) for the latter. Fix x_0 and suppose we wish to estimate $m(x_0)$. The kernel and nearest neighbor estimates are defined as follows. Let K be a nonnegative even density function.

Kernel Estimates Let h_{ker} be a bandwidth depending on n. Then the kernel estimate is

$$\hat{m}_{ker}(x_0) = \frac{\sum_{i=1}^{n} y_i K(\frac{x_i - x_0}{h_{ker}})}{\sum_{i=1}^{n} K(\frac{x_i - x_0}{h_{ker}})}. \tag{1}$$

Nearest Neighbor Estimates Let $k = k(n)$ be a sequence of positive integers, and let R_n be the Euclidean distance between x_0 and its kth nearest neighbor. Then the nearest neighbor estimate is

$$\hat{m}_{kNN}(x_0) = \frac{\sum_{i=1}^{n} y_i K(\frac{x_i - x_0}{R_n})}{\sum_{i=1}^{n} K(\frac{x_i - x_0}{R_n})}. \tag{2}$$

Under differentiability conditions on the marginal density f_x, Mack has shown that the asyptotically optimal versions of the kernel and nearest neighbor estimates have the same behavior. Let $m^{(j)}$ and $f_x^{(j)}$ denote the jth derivative of m and f_x respectively. If $c_K = \int K^2(x)dx$ and $d_K = \int x^2 K(x)dx$, remembering that K is symmetric, the kernel estimate has bias

$$bias_{ker} = h_{ker}^2 d_K \frac{m^{(2)}(x_0) f_x(x_0) + 2m^{(1)}(x_0) f_x^{(1)}(x_0)}{2 f_x(x_0)} + o(h_{ker}^2) \tag{3}$$

and variance

$$var_{ker} = c_K Var(Y \mid X = x_0)/(n h_{ker} f_x(x_0)) + o((n h_{ker})^{-1}). \tag{4}$$

There is obviously a bias versus variance tradeoff here, so that if one wants to achieve the minimum mean squared error, the optimal bandwidth is $h_{ker} \sim n^{-1/5}$ and the optimal mean squared error is of order $O(n^{-4/5})$. The formulae for bias and variance of the kth nearest neighbor estimate are the same as in (3) and (4) if one substitutes $2f_x(x_0)nh_{ker}$ for k.

Let F denote the distribution function of X, and let F_n denote the empirical distribution of the sample from X. Let h_{nn} be a bandwidth tending to zero. The estimate proposed by Yang (1981) and studied by Stute (1984) is

$$\hat{m}_{nn}(x_0) = (nh_{nn})^{-1} \sum_{i=1}^{n} y_i K\left(\frac{F_n(x_i) - F_n(x_0)}{h_{nn}}\right). \tag{5}$$

The nearest neighbor estimate defines neighbors in terms of the Euclidean norm, which in this case is just absolute difference. The estimate (5) is also a nearest neighbor estimate, but now neighbors are defined in terms of distance based on empirical distribution function. This makes for computational efficiency if the uniform kernel is used. A direct application of (5) would result in $O(n^2h)$ operations, but using updating as the window moves over the span of the x's results in $O(n)$ operations. Other smooth kernels can be computed efficiently by iterated smoothing, i.e., higher order convolution of the uniform kernel. Another possible device is the Fast Fourier transform (Härdle, 1987). Since the difference between (2) and (5) is that (5) picks its neighbors symmetrically, we call it a symmetrized nearest neighbor estimate. Note that $\hat{m}_{kNN}$ always averages over a symmetric neighborhood in the x-space, but may have an asymmetric distribution of x points in this neighborhood. By contrast, $\hat{m}_{nn}$ always averages over the same amount of points left and right of x_0, but may in effect average over an asymmetric neighborhood in the x-space. The estimate $\hat{m}_{nn}$ has an intriguing relationship with the k-NN estimator used by Friedman (1986). The variable span smoother proposed by Friedman uses the same type of neighborhood as does $\hat{m}_{nn}$ and is used as an elementary building block for ACE, see Breiman and Friedman (1985). The estimate (5) also looks appealingly like a kernel regression estimate of Y against not X but rather $F_n(X)$. Define

$$\bar{m}_{nn}(x_0) = h_{nn}^{-1} \int m(x) K\left(\frac{F(x) - F(x_0)}{h_{nn}}\right) dx. \tag{6}$$

Then Stute shows that as as $n \to \infty$, $h_{nn} \to 0$ and $nh_{nn}^3 \to \infty$,

$$(nh_{nn})^{1/2}(\hat{m}_{nn}(x_0) - \bar{m}_{nn}(x_0)) \Rightarrow Normal(0, c_K Var(Y \mid X = x_0)). \tag{7}$$

This has the form (4) as long as $h_{nn} = h_{ker} f_x(x_0)$. With this choice of h_{nn}, Stute's estimate has the same limit properties as a kernel or ordinary nearest

neighbor estimate as long as its bias term satisfies (3). Stute shows that the bias is of order $O(h_{snn}^2)$, although he does not give an asymptotic formulae. It is in fact easy to show that the bias satisfies to order $o(h_{snn}^2)$

$$(8) \qquad bias_{snn} = h_{snn}^2 d_K \frac{m^{(2)}(x_0) f_x(x_0) - m^{(1)}(x_0) f_x^{(1)}(x_0)}{2 f_x^3(x_0)}.$$

Comparison of (3) and (8) shows that even when the variances of all three estimates are the same (the case $h_{snn} = h_{ker} f_x(x_0)$), the bias properties differ unless

$$m^{(1)}(x_0) f_x^{(1)}(x_0) = 0.$$

Otherwise, the optimal choice of bandwidth for the kernel and ordinary nearest neighbor estimates will lead to a different mean squared error than what obtains for the symmetrized nearest neighbor estimate.

The preceeding discussion presumed that we are interested in estimating the regression function only at the point x_0 and that bandwidth was chosen locally so as to minimize asymptotic mean squared error. In practice, one is usually interested in the regression curve over an interval, and the bandwidth is chosen globally, see for example Härdle, Hall and Marron (1988). Inspection of (3), (4) and (8) shows the usual tradeoff between kernel and nearest neighbor estimates: in the tails of the distribution of x, the former are more variable but less biased.

The symmetrized nearest neighbor estimate is a kernel estimate based on transforming the x data by F_n. Other transformations are possible, e.g., $log(x)$. In general, if we transform by $w = G(x)$, if $m_*(w) = m(x)$ and w has density f_w, then the bias and variance properties of the resulting kernel estimate are given by (3)-(4) in m_* and f_w, the translation to f_x and m being immediate by the chain rule.

An Example

For illustrative purposes we use a large data set (n=7125) of the relationship of Y = expenditure for potatoes versus X = net income of British households (in tenth of a pence) in 1973. The data come from the *Family Expenditure Survey, Annual Base Tapes 1968-1983*, Department of Employment, Statistics Division, Her Majesty's Stationary Office, London, and were made available by the ESRC Data Archive at the University of Essex. See Härdle (1988, Chapter 1) for a discussion . For these data, we used the quartic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1).$$

We computed the ordinary kernel estimate (1) and the symmetrized nearest neighbor estimate (5), the bandwidths being selected by crossvalidation, see Härdle and Marron (1985). The crossvalidated bandwidths were $h_{ker} = 0.25$ on the scale (0,3) of Figure 1 and $h_{nn} = 0.15$ on the F_n scale. The resulting regression curves are plotted in Figure 1. The two curves are similar for $x \leq 1$, which is where most of the data lie. There is a sharp discrepancy for larger values of x, the kernel estimate showing evidence of a bimodal relationship and the symmetrized nearest neighbor estimate indicating either an asymptote or even a slight decrease as income rises. In the context, the latter seems to make more sense economically and looks quite similar to to curve in Hildenbrand and Hildenbrand 1986). Statistically, it is in this range of the data that the density f_x takes on small values, which is exactly when we expect the biggest differences in the estimates, i.e., the kernel estimate should be more variable but less biased.

References

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-619.

Friedman, J. (1986). A variable span smoother. Department of Statistics Technical Report LCS5, Stanford University.

Härdle, W. (1987). Resistant smoothing using the Fast Fourier transform, AS 222. *Applied Statistics*, 36, 104-111.

Härdle, W. (1988). *Applied Nonparametric Regression.* Book to appear.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters away from their optimum? (with discussion). *Journal of the American Statistical Association*, to appear.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13, 1465-1481.

Hildenbrand, K. and Hildenbrand, W. (1986). On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics*, W. Hildebrand and A. Mas-Colell, editors. North Holland.

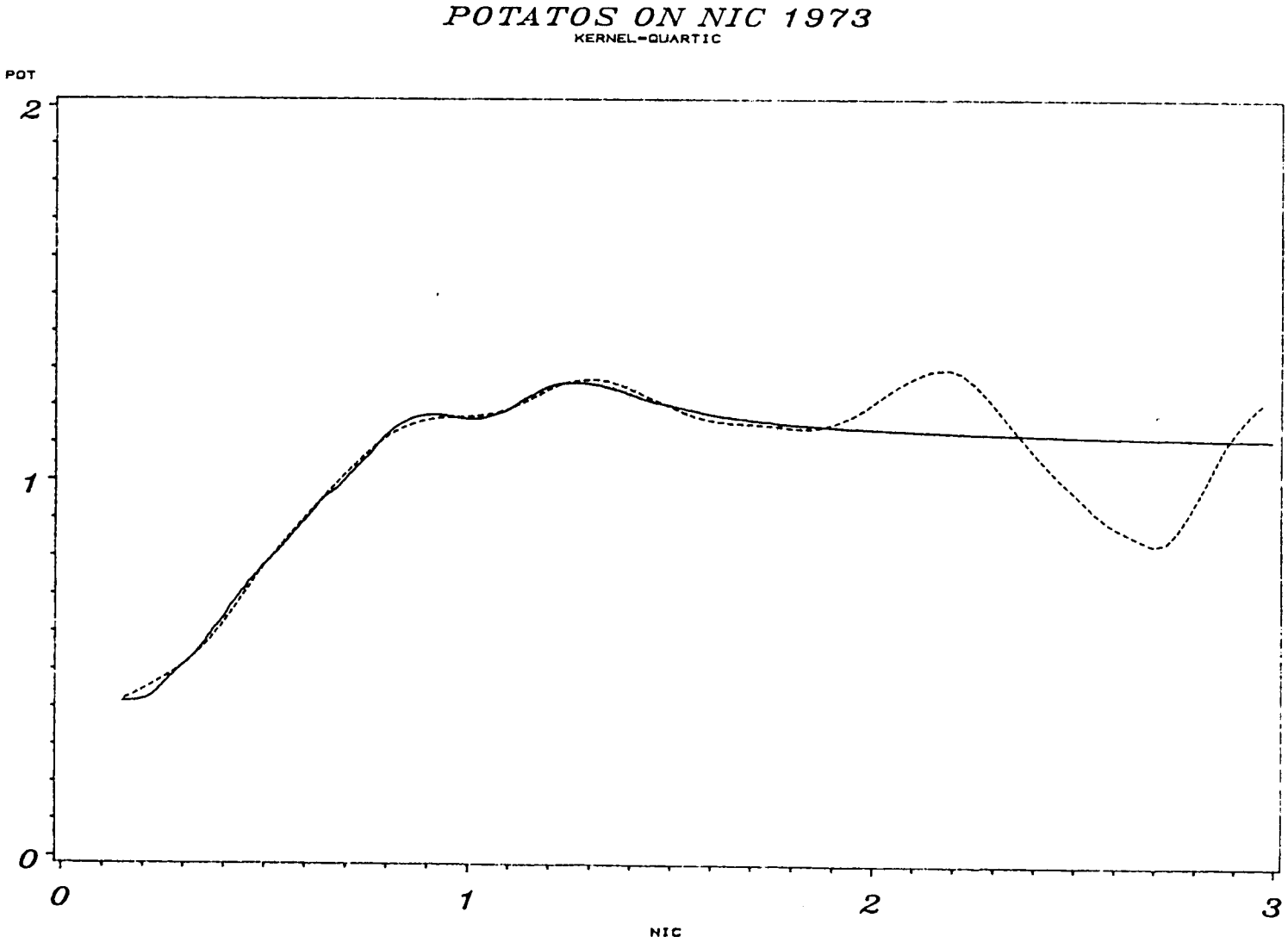
Mack, Y. P. (1981). Local properties of k-NN regression estimates. *SIAM J. Alg. Disc. Meth.*, 2, 311-323.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Annals of Statistics*, 12, 917-926.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā, Series A*, 26, 359-372.

Yang, S. S. (1981). Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function. *Journal of the American Statistical Association*, 76, 658-662.


POTATOS ON NIC 1973
KERNEL-QUARTIC
POT
2
1
0
0
1
2
3
NIC